importing-data-frames

```
# Load dataset from CSV and preview first few rows
birthweight <- read.csv("birthweight.csv", stringsAsFactors = FALSE)
head(birthweight)
```

	ID	birth.date	location	length	birthweight	head.circumference	weeks.gestat
	<int>	<chr>	<chr>	<int>	<dbl>	<int>	<i
1	1107	1/25/1967	General	52	3.23	36	
2	697	2/6/1967	Silver Hill	48	3.03	35	
3	1683	2/14/1967	Silver Hill	53	3.35	33	
4	27	3/9/1967	Silver Hill	53	3.55	37	
5	1522	3/13/1967	Memorial	50	2.74	33	
6	569	3/23/1967	Memorial	50	2.51	35	

6 rows | 1-9 of 19 columns

```
# Calculate range of paternal ages
range_paternal <- max(birthweight$paternal.age, na.rm = TRUE) - min(birthweight$pa
ternal.age, na.rm = TRUE)
cat("Range of paternal ages:", range_paternal, "\n")
```

```
## Range of paternal ages: 27
```

```
# Convert smoker column from "yes"/"no" strings to logical TRUE/FALSE
# Step 1: Inspect conversion (not strictly necessary but left for clarity)
as.logical(birthweight$low.birthweight)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [25] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [37]  TRUE  TRUE FALSE FALSE FALSE FALSE
```

```
as.logical(birthweight$smoker)
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
# Step 2: Check string values
birthweight$smoker == "yes"
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## [13]  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
## [25] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
## [37]  TRUE FALSE FALSE  TRUE  TRUE FALSE
```

```
# Step 3: Convert to TRUE/FALSE
birthweight$smoker <- (birthweight$smoker == "yes")
```

```
# Run a chi-squared test between geriatric pregnancy status and low birthweight
?chisq.test
chisq.test(birthweight$geriatric.pregnancy, birthweight$low.birthweight)
```

```
## Warning in chisq.test(birthweight$geriatric.pregnancy,
## birthweight$low.birthweight): Chi-squared approximation may be incorrect
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  birthweight$geriatric.pregnancy and birthweight$low.birthweight
## X-squared = 2.7398e-31, df = 1, p-value = 1
```

```
# Compare mean birthweight between geriatric and non-geriatric pregnancies
mean(birthweight$birthweight[birthweight$geriatric.pregnancy])
```

```
## [1] 3.1125
```

```
# the ! character is used for negation
mean(birthweight$birthweight[!birthweight$geriatric.pregnancy])
```

```
## [1] 3.333947
```

```
# Calculate mean and standard deviation of paternal age
mean(birthweight$paternal.age, na.rm = TRUE)
```

```
## [1] 28.76316
```

```
sd(birthweight$paternal.age, na.rm = TRUE)
```

```
## [1] 7.061254
```

```
# Split birth.date column into separate month/day/year columns using strsplit
?strsplit
strsplit(birthweight$birth.date, split = "/")
```

```
## [[1]]
## [1] "1"    "25"   "1967"
## 
## [[2]]
## [1] "2"    "6"    "1967"
## 
## [[3]]
## [1] "2"    "14"   "1967"
## 
## [[4]]
## [1] "3"    "9"    "1967"
## 
## [[5]]
## [1] "3"    "13"   "1967"
## 
## [[6]]
## [1] "3"    "23"   "1967"
## 
## [[7]]
## [1] "4"    "23"   "1967"
## 
## [[8]]
## [1] "5"    "5"    "1967"
## 
## [[9]]
## [1] "6"    "4"    "1967"
## 
## [[10]]
## [1] "6"    "7"    "1967"
## 
## [[11]]
## [1] "6"    "14"   "1967"
## 
## [[12]]
## [1] "6"    "20"   "1967"
## 
## [[13]]
## [1] "6"    "25"   "1967"
## 
## [[14]]
## [1] "7"    "12"   "1967"
## 
## [[15]]
## [1] "7"    "13"   "1967"
## 
## [[16]]
## [1] "9"    "7"    "1967"
## 
## [[17]]
## [1] "10"   "7"    "1967"
## 
## [[18]]
## [1] "10"   "19"   "1967"
```

```
##
## [[19]]
## [1] "11"   "1"    "1967"
##
## [[20]]
## [1] "12"   "7"    "1967"
##
## [[21]]
## [1] "12"   "14"   "1967"
##
## [[22]]
## [1] "1"    "8"    "1968"
##
## [[23]]
## [1] "1"    "10"   "1968"
##
## [[24]]
## [1] "1"    "21"   "1968"
##
## [[25]]
## [1] "2"    "2"    "1968"
##
## [[26]]
## [1] "2"    "16"   "1968"
##
## [[27]]
## [1] "2"    "22"   "1968"
##
## [[28]]
## [1] "4"    "2"    "1968"
##
## [[29]]
## [1] "4"    "24"   "1968"
##
## [[30]]
## [1] "4"    "25"   "1968"
##
## [[31]]
## [1] "6"    "19"   "1968"
##
## [[32]]
## [1] "7"    "18"   "1968"
##
## [[33]]
## [1] "7"    "24"   "1968"
##
## [[34]]
## [1] "8"    "12"   "1968"
##
## [[35]]
## [1] "8"    "17"   "1968"
##
## [[36]]
## [1] "9"    "7"    "1968"
```

```
##
## [[37]]
## [1] "9"    "16"   "1968"
##
## [[38]]
## [1] "9"    "27"   "1968"
##
## [[39]]
## [1] "10"   "9"    "1968"
##
## [[40]]
## [1] "10"   "25"   "1968"
##
## [[41]]
## [1] "12"   "11"   "1968"
##
## [[42]]
## [1] "12"   "19"   "1968"
```

```
# custom function takes a vector of dates and returns a data frame with columns da
y, month, and year
split_MMDDYYYY <- function(date_vector){
  date_list = lapply(seq(1:3), function(i){
    as.integer(sapply(strsplit(date_vector, split = "/"), '[[', i))
  })
  names(date_list) = c("month", "day", "year")
  as.data.frame(do.call("cbind", date_list))
}

# Apply date-splitting function and merge results with main data frame
split_MMDDYYYY(birthweight$birth.date)
```

month	day	year
<int>	<int>	<int>
1	25	1967
2	6	1967
2	14	1967
3	9	1967
3	13	1967
3	23	1967
4	23	1967
5	5	1967
6	4	1967
6	7	1967

1-10 of 42 rows Previous **1** 2 3 4 5 Next

```
birthweight <- cbind(birthweight, split_MMDDYYYY(birthweight$birth.date))
```

```
# Calculate mean maternal age
mean_maternal_age <- mean(birthweight$maternal.age, na.rm = TRUE)
cat("Mean maternal age:", mean_maternal_age, "\n")
```

```
## Mean maternal age: 25.54762
```

```
# Find the index of the mother who smoked the most
heaviest_smoker_index <- which.max(birthweight$maternal.cigarettes)

# Retrieve her age
age_heaviest_smoker <- birthweight$maternal.age[heaviest_smoker_index]
cat("Age of mother who smoked the most:", age_heaviest_smoker, "\n")
```

```
## Age of mother who smoked the most: 37
```

```
# Compare pre-pregnancy weight between mothers of low and normal birthweight babie
s

# Calculate group-wise means
mean_lbw <- mean(birthweight$maternal.prepregnant.weight[birthweight$low.birthweig
ht == 1], na.rm = TRUE)
mean_non_lbw <- mean(birthweight$maternal.prepregnant.weight[birthweight$low.birth
weight == 0], na.rm = TRUE)

# Output group means
cat("Mean pre-pregnant weight for mothers of LOW birthweight babies: ", mean_lbw,
"\n",
    "Mean pre-pregnant weight for mothers of NORMAL birthweight babies: ", mean_no
n_lbw, "\n")
```

```
## Mean pre-pregnant weight for mothers of LOW birthweight babies:  51.33333
##  Mean pre-pregnant weight for mothers of NORMAL birthweight babies:  58.52778
```

```r
# Interpret result
if (!is.na(mean_lbw) && !is.na(mean_non_lbw)) {
  if (mean_lbw > mean_non_lbw) {
    cat("➡️ Pre-pregnant weight is HIGHER among low birthweight group.\n")
  } else if (mean_lbw < mean_non_lbw) {
    cat("Pre-pregnant weight is LOWER among low birthweight group.\n")
  } else {
    cat("The mean pre-pregnant weight is the SAME in both groups.\n")
  }
} else {
  cat("Cannot compare means — NA values still exist.\n")
}
```

```
## Pre-pregnant weight is LOWER among low birthweight group.
```